



Student Performance Assessment Using Clustering Techniques

Noel Varela¹(✉), Edgardo Sánchez Montero¹, Carmen Vásquez²,
Jesús García Guilianny³, Carlos Vargas Mercado⁴,
Nataly Orellano Llinas⁵, Karina Batista Zea⁴, and Pablo Palencia⁵

¹ Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
{nvarela2, esanchez2}@cuc.edu.co

² Universidad Nacional Experimental Politécnica “Antonio José de Sucre”,
Barquisimeto, Venezuela
cvasquez@unexpo.edu.ve

³ Universidad Simón Bolívar, Barranquilla, Colombia
jesus.garcia@unisimonbolivar.edu.co

⁴ Corporación Universitaria Latinoamericana, Barraquilla, Colombia
carlosvargasmmercado0103@gmail.com,
kbatistazea@hotmail.com

⁵ Corporación Universitaria Minuto de Dios – UNIMINUTO,
Barranquilla, Colombia
Nataly.Orellano@gmail.com,
pablo.palencia.d@uniminuto.edu.co

Abstract. The application of informatics in the university system management allows managers to count with a great amount of data which, rationally treated, can offer significant help for the student programming monitoring. This research proposes the use of clustering techniques as a useful tool of management strategy to evaluate the progression of the students’ behavior by dividing the population into homogeneous groups according to their characteristics and skills. These applications can help both the teacher and the student to improve the quality of education. The selected method is the data grouping analysis by means of fuzzy logic using the Fuzzy C-means algorithm to achieve a standard indicator called Grade, through an expert system to enable segmentation.

Keywords: Clustering · Fuzzy C-means algorithm · Fuzzy logic · Expert system

1 Introduction

In Latin America, higher education public institutions currently face the challenge of improving their academic quality with scarce financial resources and, at the same time, coping with the demands of the new social and economic contexts in the global society [1, 2]. For this reason, there is an evident concern about developing processes and products at both academic and administrative levels and optimizing the use of available resources. An interest of great importance for university authorities is the student educational outcome whose study and analysis require the use of adequate tools for generating indicators that guide the decision-making processes at this educational level [3].

Regarding the issues mentioned above, several authors focus on the student performance as one of the most critical and urgent problems to face, since this issue is usually showing low academic results of students at Latin America universities. The impact of this problem is widely known in relation to the high drop-out rates, high repetition rates, high number of students with a lag in their studies, low grade averages, and low licensing rates [4–6].

In education, grading is the process of applying standardized measures of different levels to assess the performance of the students in a course. As a result, Grading can be used by potential employers or educational institutions to evaluate and compare the applicant's knowledge. On the other hand, data grouping is the process of extracting previously unknown, valid, useful, and positionally hidden patterns of large data sets. The main objective of clustering is to divide the students into homogeneous groups according to their characteristics and skills [7]. These applications can help both the teacher and the student to improve the quality of education. This research applies the cluster analysis to segment students into groups according to their characteristics [8].

Data Clustering is a statistical technique for data analysis without supervision. It is used for classification of homogeneous groups to discover patterns and hidden relationships that help make decisions in a quick and efficient way. By the use of this technique, a large data set is segmented into subsets called clusters (groups) [9–11]. Each cluster is a collection of similar objects, gathered into the same group, and different to objects in other clusters. Concepts of Fuzzy Logic are defined to perform the clustering, which finally leads to the determination of the Grade indicator to obtain the student performance.

2 Theoretical Review

2.1 Data Grouping (Clustering)

A clustering algorithm organizes items into groups based on similarity criteria. The Fuzzy C-means is a clustering algorithm where each item can belong to more than one group - hence the word Fuzzy - where the membership degree for each element is given by a probability distribution on the clusters [12, 13].

2.1.1 Fuzzy C-Means Algorithm (FCM)

FCM is a clustering algorithm developed by Dunn and subsequently improved by Bezdek. It is useful when the required number of clusters is predetermined. Therefore, the algorithm tries to put each data points in one of the clusters. What makes FCM different is that it does not decide the allocation of a data point to a certain group, instead, it calculates the probability of that point to belong to that cluster (degree of membership). Therefore, depending on the required accuracy of clustering, appropriate tolerance measures can be used. Given that the absolute composition is not calculated, FCM can be extremely fast because the number of iterations needed to achieve a specific grouping corresponds to the accuracy required [14, 15].

a.1 – Iterations. In each iteration of FCM algorithm, the following objective function J is minimized, Eq. 1 [16]:

$$J = \sum_{i=1}^N \sum_{j=1}^C \delta_{ij} \|x_i - c_j\|^2 \quad (1)$$

Where N is the number of data points, C is the number of needed clusters, c_j is the vector of centers for the cluster j , and δ_{ij} is the degree of membership to the i -th data point x_i on the cluster j . The norm $\|x_i - c_j\|$ measures the similarity (or closeness) of the data point x_i to the vector of centers c_j in the cluster j . Note that, in each iteration, the algorithm maintains a vector of centers for each cluster. These data points are calculated as their own weighted average, where the weights are given by the degrees of membership.

a.2 - Degrees of Membership. For a data point x_i , the degree of membership in the cluster j is calculated as follows, Eq. 2 [17]:

$$\delta_{ij} = \frac{1}{\sum_{k=1}^C \left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right]^{\frac{2}{m-1}}} \quad (2)$$

Where, m is the fuzziness coefficient and the vector of centers c_j is calculated as follows, Eq. 3 [17]:

$$c_j = \frac{\sum_{i=1}^N \delta_{ij}^m \cdot x_i}{\sum_{i=1}^N \delta_{ij}^m} \quad (3)$$

In the previous Eq. (3), δ_{ij} is the value of the degree of membership calculated in the previous iteration. Note that, at the beginning of the algorithm, the degree of membership for the data point i to the cluster j began with a random value θ_{ij} , $0 < \theta_{ij} < 1$, such as the Eq. 4 [16]:

$$\sum_j^C \delta_{ij} = 1 \quad (4)$$

a.3 - Fuzziness Coefficient. In Eqs. (2) and (3), the fuzziness coefficient m , with $1 < m < \infty$ measures the required clustering tolerance. This value determines how many clusters may overlap with each other. The higher the value of m , the greater the overlap between clusters. In other words, the higher the fuzziness coefficient used by the algorithm, a greater number of data points fall within a “fuzzy” band where the degree of membership is neither 0 nor 1 but will be somewhere in the middle [12, 18].

a.4 - Termination Condition. The required accuracy of the membership degree determines the number of iterations made by the FCM algorithm. This measure of accuracy is calculated using the degree of membership of an iteration to the next, taking the largest of these values in all the data points considering all groups. If the measure of

accuracy between iteration k and $k + 1$ is represented with ε , its value is calculated in the following way, Eq. 5 [18]:

$$\varepsilon = \Delta_i^N \cdot \Delta_j^C \cdot |\delta_{ij}^{k+1} - \delta_{ij}^k| \quad (5)$$

Where, δ^k_{ij} and δ^{k+1}_{ij} are respectively the degrees of membership in the iteration k and $k + 1$, and the operator δ , when supplied with a vector of values, returns the largest value of that vector.

3 Method

3.1 Database

The data used in this study was obtained from a Moodle 3.1 course consisting of 4,115 university students of Industrial Engineering from a private university in Colombia. The study was conducted during the 2017–2018 academic term. The research is carried out on a mandatory ninth-year subject matter.

3.2 Methods

3.2.1 Fuzzy Logic

Fuzzy logic manages the imprecision and uncertainty in a natural way, where it is possible, to provide a representation of knowledge with human orientation [4]. In recent years, there are many research studies where fuzzy logic, neural networks, classic neural networks, and the fuzzy neural networks have been employed on student modeling systems. This research applies the Fuzzy C-means Clustering algorithm for the automatic generation of the membership function. The Rule Based Fuzzy Expert System is also applied to automatically convert the crisp data into a fuzzy set and calculate the total score of a student in the exams of the first term (parc-1), second term (parc-2), and third term (parc-3). This type of knowledge in the management system can improve policies and strategies for enhancing the system quality [19, 20].

(i). Crisp value: The crisp value is the grade obtained by the student in the exams constituting the crisp input. (ii). Fuzzification: Fuzzification means that the crisp value (student's grade) becomes the fuzzy input value with the help of the proper membership function. (iii). Inference Mechanism: Defines the different types of fuzzy rules ("If/Then Rule") to assess the academic performance of students. (iv). Fuzzy Output: Determines an output value of the membership function for each active rule ("If/Then Rule"). (v). Defuzzification (Performances): Defuzzification means calculating the final result (performance value) with the help of the proper defuzzification method. In this research, Center of Area (COA) is used for defuzzification (performance assessment) [21, 22].

The membership function is a graphical representation of the magnitude of participation of each entry. A chart that defines how the value of membership between 0 and 1 is assigned to each point in the input space. The input space is often referred to as the universe of discourse or universal set that contains all possible elements of

particular interest. A “weight” is associated with each of the processed entries, the functional overlap between entries is defined, and finally, determines an output response. The rules use the input membership values as weighting factors to determine their influence on the fuzzy output sets of the conclusion in the final output. Once the functions are inferred, they are scaled and combined, and they are then defuzzified in a crisp output that activates the system [21, 23].

3.2.2 Fuzzification

Fuzzification consists on the process of transforming crisp values into degrees of membership for linguistic terms of fuzzy sets. The membership function is used to associate a score to each linguistic term. It also refers to the transformation of an objective term into a fuzzy concept. All this activity is performed by Matlab [5] assisted by the Toolbox of Fuzzy Logic, through the FCM function, which format is: [centers, U,objFun] = fcm(data,Nc,options) [24].

Tickets for this function are [23]:

- Data: Matrix with input data that should be grouped, specified as an array with Nd rows, where Nd is the number of data points. The number of columns of data is equal to the dimensionality of the data.
- Nc: Number of clusters chosen by the user.
- Options(1): Exponent of the matrix of fuzzy partition U, specified as a scalar greater than 1.0. This option controls the amount of fuzzy overlap between the groups, with higher values indicating a greater degree of overlap (default = 2).
- Options(2): Maximum number of iterations (default = 100).
- Options(3): Difference between variations of desired centroid (default = $1e-5$).
- Options(4): Shows iterations (default = 1).

While the outputs are [24]:

- Center: Coordinates of the Centers of final clusters, returned as a matrix with Nc rows that contain the coordinates of each cluster. The number of columns in the centers is equal to the dimensionality of the data to be grouped together.
- U: Matrix of diffuse partition, returned as an array with Nc rows and Nd columns. The item $U(i, j)$ indicates the degree of membership of the data point j in the cluster i . For a given data point, the sum of the membership values for all groups is one.
- objFun: Values of the objective function for each iteration, returned as a vector.

4 Results

The data make up a numeric matrix (data) of dimension $n \times 3$ where $n = 4115$ is the number of students assessed, and the first, second, and third columns show values obtained in parc-1, parc-2, and parc-3 (Fig. 1).

83	85	5
71	46	45
64	43	24

Fig. 1. Matrix of average grades in the tests 1 to 3.

The number of chosen clusters N_c is 5 [Very High (VH), High (H), Average (A), Low (L), Very Low (VL)] with no other option, so the values of default are taken (2, 100, $1e-5$, 1). As output, the matrix centers is obtained (dimension 5×3 , Fig. 2) with the coordinates of the centers of clusters and the UT matrix (dimension $n \times 5$, Fig. 3) where the elements of each column belong to each of the five groups:

67.2672	83.5736	52.7654
63.7560	35.5156	41.5214
82.0763	89.7203	78.8954
58.0252	38.7127	23.6533
81.7372	62.5113	76.8585

Fig. 2. Matrix centers for data under study.

0.2895	0.1948	0.1365	0.2658	0.1321
0.0298	0.8384	0.0165	0.0838	0.0354
0.0294	0.1969	0.0122	0.7446	0.0188
0.1042	0.0383	0.6323	0.0284	0.1960
0.0891	0.0891	0.1185	0.0475	0.6525

Fig. 3. UT matrix for data under study.

From this point, the position of the maximum of elements of each row is determined assigning it (position index) to an element in a sixth column, as shown in Fig. 4. Subsequently, each student is grouped to one of the five categories (from left to right, VH: Very High, H: High, A: Average, L: Low; and VL: Very Low) [25].

0.1374	0.1327	0.2935	0.1947	0.2445
0.0175	0.0354	0.0363	0.8324	0.0833
0.0133	0.0188	0.0286	0.1959	0.7445
0.6324	0.1960	0.1042	0.0393	0.0283
0.1178	0.6525	0.0884	0.0941	0.0469

Fig. 4. Matrix with position index for the studied data.

After the simple observation of the maximum per row, the first student is in the category A, the second one in the L, the third one in the VL, the fourth one in the VH, etc. In percentage terms, in accordance with this technique, the clustering in the five groups shows the following format, shown in Fig. 5:

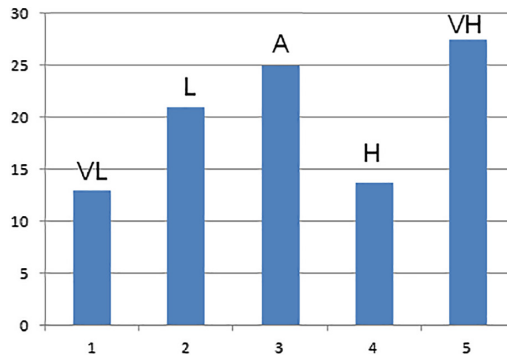


Fig. 5. Clustering of the studied data.

When the inference has ended, it is necessary to calculate a single value to represent the result. This process is called defuzzification and can be achieved by different methods. A common method is the data defuzzification in a crisp output achieved through the combination of results from the inference process and then calculating the “fuzzy centroid” of the area. The weighted strengths of each membership function of output is multiplied by their respective central points of the membership function and are added together. Finally, this area is divided by the sum of the strengths of the weighted membership function and the result is taken as the crisp output [22, 23].

Fuzzy Rule Base (Knowledge Base): The Fuzzy if-then rules and fuzzy reasoning are the backbone of the fuzzy expert systems, which are the most important modeling tools based on the fuzzy sets theory. The rule base is characterized in the form of if-then rules where the antecedents and consequents involve linguistic variables. This research considers very high, high, medium, low, and very low as linguistic variables. The collection of these rules constitutes the basis of rules for the fuzzy logic system. In this fuzzy expert system based on rules, the following rules have been used to find the knowledge base [22]:

1. If the student belongs to very high, then ($Z_1 = a_1 + b_1X_1 + c_1Y_1$)
2. If the student belongs to high, then ($Z_2 = a_2 + b_2X_2 + c_2Y_2$)
3. If the student belongs to the middle, then ($Z_3 = a_3 + b_3X_3 + c_3Y_3$)
4. If the student belongs to low, then ($Z_4 = a_4 + b_4X_4 + c_4Y_5$)
5. If the student belongs to very low, then ($Z_5 = a_5 + b_5X_5 + c_5Y_5$)

Where X_i and Y_i are the grades that the students grouped in cluster i obtained in the parc-1 and parc-2 tests, respectively, Z_i corresponding to parc-1. a_1, \dots, a_5 , b_1, \dots, b_5 and $c_1, c_5 \dots$ are constants to be determined by the method of regression analysis model.

Inference engine (Decision-Making Logic): Using the proper procedure, the true value for the antecedent of each rule is calculated and applied to the consequent part of each rule. In this case, the model of linear regression analysis for decision-making has been used. The result is a fuzzy subset to be assigned to each output variable for each rule. Once again, using a proper composition procedure, all fuzzy subsets to be assigned to each output variable are combined to form a single fuzzy subset for each output variable.

Defuzzification Interface: Defuzzification means converting the fuzzy output into crisp output. The defuzzification height was used as a technique to convert the fuzzy output into crisp output (students performance value). The defuzzification formula (Takagi-Sugeno-Kang Model) is the following, Eq. 6 [25]:

$$Y = \frac{\mu_{VH}(x, y) \cdot Z_1 + \mu_H \cdot Z_2 + \mu_A(x, y) \cdot Z_3 + \mu_L(x, y) \cdot Z_4 + \mu_{VL}(x, y) \cdot Z_5}{\mu_{VH}(x, y) + \mu_H(x, y) + \mu_A(x, y) + \mu_L(x, y) + \mu_{VL}(x, y)} \quad (6)$$

With the help of this equation, the fuzzy output can be converted into crisp output (student performance value) getting the standard indicator called Grade. The result for the first records is shown in Table 1:

Table 1. Grade indicator for the first records.

p1	p2	p3	VH	H	A	L	VL	Grade
84	84	5	0.2935	0.1947	0.1365	0.2432	0.1320	71.0740
70	44	45	0.0363	0.8284	0.0165	0.0833	0.0354	50.7925
65	44	25	0.0286	0.1959	0.0122	0.7445	0.0188	34.1655
90	93	96	0.1041	0.0393	0.6323	0.0283	0.1960	69.9056

5 Conclusions

According to the results, the research presents a new way of grouping students in terms of their academic behavior, which is more significant than using a traditional average. So, this fully developed pedagogical tool for planning according to a classification from an internationally recognized indicator is made available to teachers and administrators.

References

1. Van Dyke, T.P., Prybutok, V.R., Kappelman, L.A.: Cautions on the use of the SERVQUAL measure to ASSESS the quality of information systems services. *Decis. Sci.* **30**(3), 877–891 (1999)
2. Bonerge Pineda Lezama, O., Varela Izquierdo, N., Pérez Fernández, D., Gómez Dorta, R.L., Viloria, A., Romero Marín, L.: Models of multivariate regression for labor accidents in different production sectors: comparative study. In: Tan, Y., Shi, Y., Tang, Q. (eds.) *DMBD 2018*. LNCS, vol. 10943, pp. 43–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_5

3. Izquierdo, N.V., Lezama, O.B.P., Dorta, R.G., Viloria, A., Deras, I., Hernández-Fernández, L.: Fuzzy logic applied to the performance evaluation. Honduran coffee sector case. In: Tan, Y., Shi, Y., Tang, Q. (eds.) ICSI 2018. LNCS, vol. 10942, pp. 164–173. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93818-9_16
4. Pineda Lezama, O., Gómez Dorta, R.: Techniques of multivariate statistical analysis: an application for the Honduran banking sector. *Innovare: J. Sci. Technol.* **5**(2), 61–75 (2017)
5. Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J.: Methodology for the design of a student pattern recognition tool to facilitate the teaching - learning process through knowledge data discovery (big data). In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 670–679. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_63
6. Duque Oliva, E., Finch Chaparro, C.: Measuring the perception of service quality education by students AAUCTU Duitama. *Free Criterion Magazine*, vol. 10, no. 16, January–July 2012
7. Yao, L.: The present situation and development tendency of higher education quality evaluation in Western Countries. *Priv. Educ. Res.* **3**, 12–45 (2006)
8. Bertolin, J., Leite, D.: Quality evaluation of the Brazilian higher education system: relevance, diversity, equity and effectiveness. *Qual. High. Educ.* **14**, 121–133 (2008)
9. Cronin, J., Taylor, S.: Measuring service quality: a reexamination and extension. *J. Mark.* **56**(3), 55–68 (1992)
10. Ballesteros Román, A.: Minería de Datos Educativa Aplicada a la Investigación de Patrones de Aprendizaje en Estudiante en Ciencias. Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional, México City (2012)
11. Ben Salem, S., Naouali, S., Chtourou, Z.: A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. *Comput. Electr. Eng.* **68**, 463–483 (2018). <https://doi.org/10.1016/j.compeleceng.2018.04.023>
12. Chakraborty, S., Das, S.: Simultaneous variable weighting and determining the number of clusters—a weighted Gaussian means algorithm. *Stat. Probab. Lett.* **137**, 148–156 (2018). <https://doi.org/10.1016/j.spl.2018.01.015>
13. Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M.: Inice: a new approach for identifying the number of clusters and initial cluster centres. *Inf. Sci.* (2018). <https://doi.org/10.1016/j.ins.2018.07.034>
14. Rahman, M.A., Islam, M.Z., Bossomaier, T.: ModEx and seed-detective: two novel techniques for high quality clustering by using good initial seeds in K-Means. *J. King Saud Univ. - Comput. Inf. Sci.* **27**, 113–128 (2015). <https://doi.org/10.1016/j.jksuci.2014.04.002>
15. Rahman, M.A., Islam, M.Z.: A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowl.-Based Syst.* **71**, 345–365 (2014). <https://doi.org/10.1016/j.knosys.2014.08.011>
16. Ramadas, M., Abraham, A., Kumar, S.: FSDE-forced strategy differential evolution used for data clustering. *J. King Saud Univ. - Comput. Inf. Sci.* (2016). <https://doi.org/10.1016/j.jksuci.2016.12.005>
17. Vásquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. *J. Eng. Appl. Sci.* **12**(11), 2963–2965 (2017)
18. Torres-Samuel, M., et al.: Efficiency analysis of the visibility of Latin American Universities and their impact on the ranking web. In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 235–243. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_22
19. Bandyopadhyay, S., Maulik, U.: Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognit.* **35**, 1197–1208 (2002)

20. Tam, H., Ng, S., Lui, A.K., Leung, M.: Improved activation schema on automatic clustering using differential evolution algorithm. In: IEEE Congress on Evolutionary Computing, pp. 1749–1756 (2017). <https://doi.org/10.1109/CEC.2017.7969513>
21. Kuo, R., Suryani Erma, E., Kuo, R.: Automatic clustering combining differential evolution algorithm and k -means algorithm. In: Lin, Y.K., Tsao, Y.C., Lin, S.W. (eds.) Proceedings of the Institute of Industrial Engineers Asian Conference 2013, pp. 1207–1215. Springer, Singapore (2013). https://doi.org/10.1007/978-981-4451-98-7_143
22. Piotrowski, A.P.: Review of differential evolution population size. *Swarm Evol. Comput.* **32**, 1–24 (2017). <https://doi.org/10.1016/j.swevo.2016.05.003>
23. Kaya, I.: A genetic algorithm approach to determine the sample size for attribute control charts. *Inf. Sci. (NY)* **179**, 1552–1566 (2019). <https://doi.org/10.1016/j.ins.2008.09.024>
24. Dobbie, G., Sing, Y., Riddle, P., Ur, S.: Research on particle swarm optimization based clustering: a systematic review of literature and techniques. *Swarm Evol. Comput.* **17**, 1–13 (2014). <https://doi.org/10.1016/j.swevo.2014.02.001>
25. Omran, M.G.H., Engelbrecht, A.P., Salman, A.: Dynamic clustering using particle swarm optimization with application in image segmentation. *Pattern Anal. Appl.*, 332–344 (2016). <https://doi.org/10.1007/s10044-005-0015-5>